The research program of the Center for Economic Studies produces a wide range of theoretical and empirical economic analyses which serve to improve the statistical programs of the U.S. Bureau of the Census. Many of these analyses take the form of research papers. The purpose of the Discussion Papers is to circulate intermediate and final results of this research among interested readers within and outside the Census Bureau. The opinions and conclusions expressed in the papers are those of the authors and do not necessarily represent those of the U.S. Bureau of the Census. All papers are screened to ensure that they do not disclose confidential information. Persons who wish to obtain copies of papers, submit comments about the papers, or obtain general information about the series should contact Peter Zadrozny, Editor, Discussion Papers, Center for Economic Studies, Room 3442, FOB 3, U.S. Bureau of the Census, Washington, DC 20233 (301-763-2490).

THE LONGITUDINAL RESEARCH DATABASE (LRD): STATUS AND RESEARCH POSSIBILITIES

by

Robert H. McGuckin and George A. Pascoe, Jr.

CES 88-2 July 1988

Abstract

This paper discusses the development and use of the Longitudinal Research Data available at the Center for Economic Studies of the Bureau of the Census in terms of what has been accomplished thus far, what projects are currently in progress, and what plans are in place for the near future.

The major achievement to date is the construction of the database itself, which contains data for manufacturing establishments collected by the Census in 1963, 1967, 1972, 1977 and 1982, and the Annual Survey of Manufactures for non-Census years from 1973 to 1985. These data now reside in the Center's computer in a consistent format across all years. In addition, a large software development task that greatly simplifies the task of selecting subsets of the database for specific research projects is well underway. Finally, a number of powerful microcomputers have been purchased for use by researchers for their statistical analysis.

Current efforts underway at the Center include research on such policy-relevant issues as mergers and their impact on profits and production, high technology trade, import competition, plant level productivity, entry and exit, and productivity differences between large and small firms. Due to the confidentiality requirements of the Census data, most of their research is performed by Center staff and Special Sworn Employees. Under certain circumstances, the Center accepts user-written programs from outside researchers. These routines are executed by Center staff, and the resultant output is reviewed thoroughly for disclosure problems. The Center is also an active member of a task force working on methods on release "masked" or "cloned" microdata in public-use files that will protect the confidentiality of the data while at the same time provide a research tool for outside users.

The Center research program contributes directly to future research possibilities. The current batch of research projects is adding insight into the nature of the LRD database. This information is continually being incorporated into the Center's software system, thus facilitating yet more research activity. Moreover, since a good portion of the research involves linking the Longitudinal Research Data to other data files, such as the NSF/Census R&D data, the scope of the databases is continually being expanded. Furthermore, the Center is exploring the possibility of linking the demographic data collected by the Census Bureau to the LRD database.

The Longitudinal Research Database (LRD): Status and Research Possibilities

Introduction

The Longitudinal Research Database (LRD) is a large micro database of establishment level data constructed by pooling information from the Census of Manufactures (CM) and the Annual Survey of Manufactures (ASM). It is housed within the Census Bureau at the Center for Economic Studies (Center). This paper outlines the development of this database, its structure and current status, and the possibilities for its use in economic research.

The construction of the database was itself a major achievement. It contains linked data from 5 different censuses and 11 different annual surveys. There are 2,311,794 individual establishment-year records currently in the file and it is updated as new data become available. Thus, the LRD is one of the most ambitious and comprehensive data sets available for the study of manufacturing, and promises to provide an exciting and stimulating research environment for many years. At the same time, the sheer magnitude of the database, coupled with its complexity, means that researchers must take the time to fully understand the structure of the data before embarking on research.

The Center was established in 1982 to oversee the development of this database, to use the data to improve future Census Bureau data collection and reports, and to make the data available to

outside users. Center research focuses on such policy issues as mergers and their impact on profits and production, measurement of high technology trade, market concentration and international competition, plant level productivity, firm entry and exit, and productivity differences between large and small firms. Due to confidentiality requirements, most of this research is performed by the Center's staff and Special Sworn Employees. Nonetheless, progress has been made toward developing public use data sets which protect the confidentiality of individual respondents while providing easier access to the data.

The discussion is organized into four sections. We begin with some general observations on the characteristics that researchers desire in a database. In particular, we focus on the need for micro level detail to adequately examine many economic issues. This discussion provides the framework for the more specific remarks in the remainder of the paper. These remarks include a brief section outlining the origins of the LRD. The main portion of the paper details the major components of the LRD, the kinds of information included in the database, and related data sets available at the Center. Throughout, we try to describe research conducted at the Center as a way of providing concrete examples of the kinds of activity the LRD will support. We then briefly discuss access to the database and conclude with some observations intended to provide an overall assessment of the usefulness and flexibility of the LRD.

THE NEED FOR DETAIL IN A DATABASE

Economic analysis has a profound influence on data development. Researchers often approach particular problems with a well-defined theory, sophisticated econometric or statistical techniques and data that is inadequate or inappropriate for testing the theory. This situation provides the incentive for developing new data. The theory provides guidance and direction to the data development strategy.[1] Unfortunately, the need for better data often occurs when an answer to a question is required in a time frame too short to develop a new data set. Even if there is time, the costs of developing new data are often prohibitive. In these instances, the available data influences the theory and the econometric procedures used. Thus, data development also influences economic analysis.

In most research on production functions and total factor productivity, data availability dictates the estimation procedures. The absence of detailed data for specific producing units often causes researchers to use aggregate data in econometric specifications. Unfortunately, several recent papers using the LRD suggest the existence of substantial aggregation bias in estimates of productivity relationships.[2] Moreover, there are many productivity related questions that simply cannot be examined with aggregate data. John Solow (1987) argues convincingly that it is impossible to determine whether energy is a complement or substitute for other inputs using aggregate data (for example, 2-

digit manufacturing industries).

As a specific example of the need for detailed data, consider the measurement of trade flows and the technological leadership of U.S. industry. Examinations of this question have focused on the high-tech trade balance defined in terms of trade flows measured at the 3-digit industry level with high-tech industries distinguished from low-tech industries on the basis of R&D/sales ratios. Use of this procedure means that low-tech products are often included in the high-tech industry category. For instance, the office and computing equipment industry (SIC 357) includes high-tech products such as electronic computers and peripheral computing equipment. It also includes low-tech products such as adding machines and coin counters. Conclusions based on such aggregate numbers may be misleading.[3]

These examples show that the need for more detailed data is a central feature of economic research. This need cuts across all applied fields of economics. The LRD is a longitudinal micro-database consisting of individual establishment (plant) data, which provides a substantial source of detailed data.

Other Elements of Data Structure

Elements of data structure other than the level of aggregation are also important for determining the usefulness of a data set to researchers. Such elements are the aspects of the data used to classify individual records. Borrowing from the computer science

literature, the structure of a database is sometimes described in terms of data keys. The keys are the levels or classifications which can be used to categorize the data. Individual data items and variables are reported by or for these categories or keys.[4]

Although it is unlikely that any list of categories of economic data would satisfy all researchers, it is possible to list typical categories which are required for most economic research. As might be anticipated from the title of this paper, we view time as one of the most important structural characteristics. Various cross-sectional aspects of data are also regularly desired in economic research. While for some problems the plant may be the appropriate unit for analysis, the firm or enterprise affiliation of the plant is more important for other issues. The location, industry classification, and size of the plant are other important aspects of the data structure of particular interest to economic researchers. Each of these variables has been made a part of the basic key structure of the LRD. As the discussion proceeds we will highlight these structural characteristics of the LRD, but emphasize that the LRD has the flexibility to accommodate research requiring new key variables.

THE LONGITUDINAL ESTABLISHMENT DATA FILE (LED)

In the late 1970s, the Census Bureau agreed to develop a longitudinal database of individual establishments based on data collected in the CM and the ASM. The project was carried out under

the direction of Richard and Nancy Ruggles of Yale University. Initial funding was provided by the National Science Foundation, the Small Business Administration, and the Census Bureau. The product of this effort was the Longitudinal Establishment Database (LED) which contained data for establishments for 1972 to 1981.

The Center was created to facilitate access to the LED file. While the Center worked with outside researchers on several projects involving the full LED file, much of the Center's efforts at database development were focused on a balanced panel of the LED file called the Time Series file. It soon became obvious that a balanced panel strategy was inappropriate. Exits due to plant closings continually reduced the number of plants in operation from the inception of the file. Adding to the decline in the number of plants operating continuously were changes in the sample design used to collect data in non-census years. Furthermore, analysis of the births of new plants and firms had extensive direct policy and research interest.[5]

These factors led the Center to rethink its strategy in early 1987. All CM data for 1963, 1967, 1972, 1977 and 1982 and ASM data for 1973 to 1985 were grouped into a distributed database, which was termed the Longitudinal Research Database. Various indexes to the data were developed. The main consequence of this substantial undertaking is that it is now possible to generate extracts of the data using a variety of selection keys such as geographic location, industry, size, firm, etc. Panels can be selected that meet the

needs of the researcher and are not constrained to certain years. Consequently, this paper will focus on the LRD, an unbalanced panel from which various balanced and unbalanced time series may be obtained.[6]

THE LONGITUDINAL RESEARCH DATABASE

To determine if the LRD is a useful data source requires a clear understanding of what the LRD contains. Since the two principle components of the LRD are fundamentally different, we will discuss the CM first, and then contrast it with the ASM.

We want to alert the reader that our discussion concentrates on methodological issues which the researcher must be careful about when conducting research. This has a tendency to emphasize problems with the data. As already noted, the LRD has been successfully employed in a wide range of studies. The results of these studies show that the LRD is a rich data source which, although not perfect, has great potential as a research tool.

The Census of Manufactures Component

The CM is an enumeration of all establishments whose primary activity is manufacturing as classified by the Census Bureau according to the Standard Industrial Classification System (SIC). An establishment is defined as an economic unit, at a single location, where business is conducted or where services or industrial operations are performed. The basic unit of data

collection is the establishment, and accordingly, a primary data key in the LRD is the establishment.

Since 1954, the Census Bureau's mailing lists used for data collection have been obtained from the Internal Revenue Service (IRS) and the Social Security Administration (SSA). For single-establishment companies, these lists are usually sufficient for data collection purposes. However, for multi-establishment companies, the Census Bureau must request additional information, in particular, the name and address of each of the company's establishments. (An interesting byproduct of this survey is a detailed description of the firm's legal form of ownership which we discuss below.) The information from the IRS and the SSA is combined with the information from the Census Bureau survey of multi-establishment companies to form the Standard Statistical Establishment List (SSEL) which forms the basis for both the CM and the ASM.

Although the CM is a complete enumeration of all manufacturing establishments, not all establishments actually report data to the Census Bureau. Data for some establishments are obtained from other government agencies, while other data items are estimated. After the 1963 CM, it was decided to reduce the reporting burden, particularly for small companies, by making greater use of the data in the records obtained from the IRS and the SSA. Beginning in l967 some small companies were exempted from reporting their data to the Census Bureau. Instead, census-type statistics for these

establishments were developed from IRS and SSA records. The information obtained from these records includes the firm's name and address, payroll, and gross business receipts. Other statistics for these small firms are estimated. Industry average values for the unobserved variables and the sales or payroll data are used to form a ratio, and this ratio is used to estimate values for the missing data for the administrative record establishments.

In 1972 approximately 120,000 small single-establishment manufacturing firms identified as having less than 10 employees were designated as administrative record cases and were excused from filing reports. In 1977 and 1982, approximately 145,000 and 130,000 firms, respectively, were treated as administrative record cases. Although the impact of administrative record data on industry aggregates is slight, (for manufacturing as a whole, administrative record cases accounted for only 1.2 percent of the value added in 1972, 1.7 percent in 1977, and 1.3 percent in 1982), these data may be important in particular industries and for certain research topics.

The information on sales and payrolls obtained from the IRS and the SSA appear to be of high quality. Moreover, the estimation techniques for the unobserved variables work well for aggregate data. However, the methods used to estimate values for the unobserved variables in these administrative record cases may produce less useful data for microeconomic projects. The individual researcher must determine if the Census Bureau

estimation method or some alternative is more appropriate for their project.[7]

The treatment of the data collected from the approximately 220,000 remaining establishments reflects the demands of primary Census Bureau users and budget constraints. The Census Bureau's primary objective for both the CM and ASM is to publish useful and accurate current year aggregates. Consequently, the data are evaluated and edited with the accuracy of the aggregate statistics in mind. Little consideration is given to the time series or micro aspects of the data. In designing sampling plans and other collection procedures, the time and expense required to edit the data for an individual establishment is weighed against the probable effect that establishment will have on the aggregates. The result is that, during editing, data for larger establishments receive more careful consideration than the data for smaller establishments.

The Annual Survey of Manufactures Component

There are two major differences between the CM and the ASM: the number of establishments covered and the data items collected. The ASM is a sample of establishments drawn from the universe of establishments in the CM. The sample is selected during the year following each census and is used for data collection for five years. After five years a new sample is drawn. The LRD contains

data from the annual surveys for 1973 to 1985. These data were collected from four separate ASM panels: the survey samples originally drawn for 1969, 1974, 1979 and 1984. Although there is substantial overlap in the establishments present in each ASM sample, the correspondence is not perfect. Details of the sampling plan are therefore important in evaluating the possibilities of using a continuous panel of establishments. Moreover, since the sampling methodology for the ASM has changed over time, and since these changes have a significant effect on the time series which can be derived from the LRD, we describe them in some detail.

For the panels selected for 1969 and 1974, an establishment's size, industry and company affiliation determined the probability of selection. If an establishment of a multi-establishment company were included in the sample, all of the company's establishments were also required to report their data regardless of size. Thus, all firms in the ASM sample for these years were complete in the sense that all their manufacturing establishments were included.

The probability of selection for a company is related to the size of its establishments.[8] All companies with a manufacturing establishment with 250 or more employees were selected. These large companies accounted for more than two-thirds of total manufacturing employment in each of the censuses conducted from 1963 forward. Companies with smaller establishments were assigned probabilities proportional to their size.

In 1979, under severe budget pressure, the Census Bureau

adopted a new procedure for sample selection. The main change was that the probability of selection for any establishment was now solely a function of the size of the establishment itself. Its company affiliation played no part in the sample design. All establishments with 250 employees or more in the 1977 Census of Manufactures were included in the 1979 sample panel. Smaller establishments were still sampled with probabilities proportional to their size, but the plants of multi-establishment companies were not included in the sample automatically if one of the company's plants was chosen.

The 1979 panel captures about 91 percent of the total manufacturing activity (measured by total value of shipments) compared with the previous panel, but the number of sampled individual establishments was reduced significantly: from approximately 75,000 to roughly 55,000. The major effect of the change was that many small establishments of multi-establishment companies were excluded from the ASM sample. In turn, the number of companies for which complete data are collected was also substantially reduced. Approximately 5,000 companies, roughly half of the total number of companies in the ASM for which complete data would have been available under the old sampling design, reported for only a portion of their establishments under the 1979 sampling methodology. Consequently, any time series research which requires complete information on firm activities will have substantially fewer observations after 1979.

To compensate for the loss of information which resulted from the 1979 change, the 1984 ASM panel now includes all establishments of companies with value of shipments of $500 million or more in 1982.[9] As before, establishments with 250 or more employees are always included in the sample, regardless of company size, and smaller establishments are selected with a probability which is proportional to their size.

It is important to note that the sampling design has implications for analysis conducted on the basis of categorizations of the data other than national industries. Consider, for example, the establishment location information in the LRD. The location of each establishment is coded by state, SMSA,[10] county and place. A sample based on these codes permits analysis below the national level. However the selection probabilities for the ASM sample make such analysis subject to potential error. Each ASM sample provides sufficient sample points to develop estimates for national totals. But, since location is not a criterion used in determining the selection probability for a particular establishment, totals derived from aggregating the micro data may not be appropriate for sub-national levels of aggregation. For example, developing county or state-wide totals in ASM years requires reweighting the data. Similarly, irrespective of the aggregations involved, use of data from survey years requires careful consideration of the sample selection process before estimating microeconomic models. As part of the Center's software development, we plan to provide data users

with methods to account for such selection biases.

Summary

The LRD contains data for all large establishments for every year from 1972 to 1985. These data are likely to be of high quality due to the attention they receive during collection and editing. The data for smaller establishments are less reliable because they receive less attention during editing. However, the sales and payroll data for the administrative record establishments are likely to be of high quality because they are not subject to substantial response error.

The ASM samples are less likely to contain small establishments because of policies to reduce reporting burdens and costs. Also the composition of the sample of smaller establishments changes every 5 years. Establishments with more than 250 employees remain in the ASM panels over time. Even though the available time series of firms is less after 1979 than before, there are still over 6,000 complete multi-unit firms available for annual analysis and substantially more than that for census years. Taken together, these procedures imply that time series over many years will contain primarily large establishments. This applies equally to time series where the firm is the unit of observation. Finally, while the sampling procedures limit the size of continuous panels available for research, several current projects are utilizing continuous panels of over 20,000 establishments.

Data Items or Variables

From every manufacturing establishment with one or more employees the CM collects data on the establishment's inputs of labor, materials, and capital, its output of products and services, its location, and the legal form of organization of the owning firm. Associated with each establishment record is a permanent plant identification number and location. Both of these items stay with the plant from its birth until it shuts down.[11] In addition each plant is linked to a parent firm and detailed status codes allow one to trace ownership changes over time. These codes were used to identify mergers among the largest firms in each 4-digit industry for the study of conglomerate mergers by McGuckin and Andrews (1987). The same codes were used for the Lichtenberg and Siegel (1987) study of ownership changes in continuously operated plants. Lichtenberg and Siegel examined the relationship between total factor productivity growth and ownership changes using the time series panel. The McGuckin-Andrews work examined the performance of acquired lines of business in the period following their acquisition by a firm not previously operating in the same industry. This study used census year data and includes analysis of closed and opened plants. The Lichtenberg and Siegel work used yearly observations on continuously operated plants derived from the CM and the ASM.

The ASM collects the same basic measures of economic activity as the CM but, in addition collects detailed information on assets,

capital expenditures, rental payments, supplemental labor cost, retirements and depreciation (after 1976) and, in selected years, the cost of purchased services. In survey years, however, less detailed information on materials consumption and the plant's product outputs are collected. Data on individual materials consumption is not requested in survey years. Also, in survey years the value of products shipped is recorded only in terms of approximately 1,500 product classes instead of the roughly 11,000 individual products used in census years.

A detailed description of the individual data items would be too cumbersome to include here but can be found in the LED Technical Documentation (1987). A brief list of the data items gives one a good idea of the breadth of coverage. On the input side the LRD contains: total employment, number of production workers (quarterly), production worker hours (quarterly), total salaries and wages, production worker wages, other employee wages, total supplemental labor costs, cost of materials, cost of resales, cost of fuels, cost and quantity of purchased electricity, cost of contract work, cost of purchased services, building and machinery repairs, cost of purchased communications services, inventory stocks at the beginning and end of the year for finished products, work-in-process and materials, capital expenditures for building and equipment, rental payments, capital stocks of buildings and equipment at the beginning and end of the reporting year, building and machinery depreciation, retirements, rents and repairs.

Appendix A provides a complete list.

The output data include value of shipments reported for each 7-digit product in CM years and for each 5-digit product class in ASM years. Related information, such as value added, miscellaneous receipts, value of resales and receipts for contract work, are also available for each establishment.

There are two important points to keep in mind when designing research projects with the LRD. First, the reporting unit for data collection is the establishment. Second, for the most part price data are only collected in census years in the form of unit values.[12] As noted by our discussant, the units (quantity) are not always well defined. For example the 7-digit level of detail does not distinguish between a $200 10-speed bicycle and a $1000 racing bicycle.

Because the establishment is the unit of observation the various inputs used by the establishment are not allocated to the specific products produced by the establishment. In most applications and for most Census Bureau published tabulations a plant is classified by the industry which accounts for the plant's largest output. As noted, detailed information on the value of shipments and physical output of products, at the 7-digit level in census years and 5-digit level in survey years, are available for each plant. The other variables are reported at the level of the entire establishment.[13]

The second point we want to emphasize is that outside of

census years there is little in the way of price information in the LRD. This means that price series for purposes of, for example, deflation in production function estimation must be obtained from non-Census Bureau sources for annual time series analysis.

This problem was recognized early on by researchers studying total factor productivity. Fortunately the Bureau of Industrial Economics (BIE) at main Commerce published an SIC-based price series based on Bureau of Labor Statistics (BLS) data. This series has been used by several researchers working with the continuous panel.[14] While there have been a number of specific research projects using the LRD, an NSF sponsored Resources for the Future (RFF) study is developing a complete data set for research into productivity issues. Phase I of the study established the feasibility of producing a balanced panel containing detailed output, price and input data. Preliminary analysis of the information developed for selected industries was reported at the American Economic Association meetings in 1987. A second phase of this work has the goal of developing a full scale data set incorporating the methodological lessons learned in Phase I. Unfortunately budget cuts may prevent the completion of Phase II.[15]

One final point with regard to the price data available in census years: These unit value figures are obtained by dividing total product (or establishment) value of shipments by the quantity produced. They represent an "average" value for all the outputs of the establishment or product class. In this sense they better

represent the combined outputs of the plant than the BLS prices which are based on particular quality adjusted products. There has been little research on the relative usefulness of these alternative measures. We explicitly raise this point because there appears to be a tendency to deemphasize unit value collection as a way to meet budget reductions. This may be very shortsighted since it is not clear that BLS price indexes are appropriate in all cases.[16]

Related Data Files

The tendency for data availability to influence the development and testing of economic models is evident in many of the research projects undertaken at the Center and described above. Of perhaps more interest are data development efforts associated with the Center's research agenda. These efforts satisfy the needs of economic researchers with regard to particular projects, and also extend the Center's ability to support new research proposals. In this section we highlight several projects involving extensions of the LRD database which have been driven by the requirements of particular research projects. Each of these extensions involves linking the LRD to another database. Some of these efforts, like the use of BIE price index data discussed above, involve an outside database. Other examples involve specialized Census Bureau surveys.

Aside from the price indexes, most of the work of expanding

the LRD involves Census Bureau data. One exception is the McGuckin-Andrews work in which stock market premium data and other financial statistics are being linked to LRD based performance measures obtained for acquired lines of business (market share, profits, and productivity). This effort is an attempt to reconcile the disparate findings regarding the gains to takeovers found in the literature. Financial market studies show substantial gains which are not observed in accounting studies.[17]

Currently, the linking of company level data to LRD companies is being made by name matches. A similar procedure has been used to match companies reporting R&D data in the NSF-sponsored R&D survey to companies in the LRD. This latter work has resulted in several published papers based on large firms.[18] Currently the R&D and LRD linking is being extended to small firms with supplemental NSF support. Completion of this work will mean that the entire R&D survey data will be linked to the LRD. One future project which could have big payoffs would be the development of an association between Census Bureau identification numbers and CUSIP numbers or similar identifiers used to identify companies in public financial databases. Such a step would improve research possibilities at the Center. It would also help to supplement the LRD by making it possible to include the operations of firms outside manufacturing in research designs. Restricting analysis of a firm's activities to those in manufacturing is unnecessarily limiting.[19]

There are several areas in which the Center is moving to

expand the LRD's compatibility with existing Census Bureau data. One major area is foreign trade where the increasingly global nature of the economy has made it necessary to merge foreign trade data with domestic statistics. Because the foreign trade data are collected on a product basis, it is sometimes difficult to reconcile the foreign trade data with LRD data collected under the SIC system. The Center is currently heading up a task force at the Census Bureau examining the feasibility of producing trade adjusted concentration and market penetration statistics for detailed product classes (5- and 7-digit). The project includes CM, ASM, and Current Industrial Reports data. If the product codes and firm identifiers can be successfully linked, then these data can also be linked to the LRD. One of the first studies will examine the impact of foreign imports in domestic markets. In turn, research involving the linked data should help refine edit procedures and provide for adjustments in collection procedures when necessary.

Finally, a major long term interest of the Center is the exploitation of individual data collected through the population censuses and surveys. The Center has at least one project which will make use of both LRD and demographic information.[20] The Center also has recently become the repository for the relatively new Survey of Characteristics of Business Owners (CBO). This survey was first conducted in 1982 and there is hope that a new panel can be developed for 1987. It is the only Census Bureau survey which directly links the characteristics of business owners with the

characteristics of the businesses they operate. This data will greatly expand our ability to examine the nature and characteristics of entrepreneurs.

ACCESSING THE DATA

Establishment data are collected by the Census Bureau under the authority of Title 13 of the United States Code. To protect confidentiality, Title 13 and the disclosure rules and regulations of the Census Bureau prohibit the release of information that could be used to identify or closely approximate the data for an individual establishment or enterprise. In practice the Census Bureau has taken disclosure protection as a binding constraint and provided as much public information as possible within this constraint. The Census Bureau has well-defined procedures for evaluating and releasing aggregate data and tabulations, but not for microdata files. As a result, only a limited number of outside researchers working at the Census Bureau as Special Sworn Employees (such as NSF/Census research fellows and associates) have had access to the LRD.[21]

For non-Census researchers, there are many advantages in coming to the Census Bureau to undertake a research project. Working at the Census Bureau provides an opportunity for outside researchers to interact with the staff who regularly work with, collect and analyze the data which compose the LRD.

While there are important gains to conducting research at the

Census Bureau, it is not without its costs. It is often difficult for researchers to relocate to Washington. Even assuming that a researcher can spend time at the Census Bureau, the facilities for supporting outside researchers are limited at the Center. The Center does not have the staff, space or computer resources to accommodate more than a few full-time visitors at any one time. Currently it can accommodate three or four researchers and will likely be able to handle a few more in the near future. The ASA/NSF/Census fellowship program which has separate facilities adds to the number of visiting scholars who can use the LRD.

The practical considerations which make it impossible to accommodate all demands for microdata by allowing outside researchers to work at the Census Bureau as Special Sworn Employees, has led to considerable interest in the development of public use files. A public use data file will be similar to the original data file in its major structural characteristics so that the important economic relationships among variables in the file are maintained. Ideally, the public use data file would preserve the economic relationships with sufficient precision so that elasticities and other parameters of interest could be obtained directly without any need for processing by the Center.[22]

In line with the public use data concept, the provision of researchers with a mock file which they can use to debug programs written in SAS or other standard packages for execution by the Center is a way to increase the access to the LRD. For projects

involving the new and relatively clean CBO database we hope to be able to provide complete processing without the researcher having to obtain special employee status. For LRD projects, until we have developed better software for editing the data and have more experience with it, most researchers will still need to visit the Center to examine the data.[23] Nonetheless, with the use of programs debugged outside the Center, the necessary time required at the Center is reduced. This means that research costs are reduced and the Center can accommodate more LRD users.

CONCLUDING COMMENT

We began our discussion by emphasizing the need for detailed microdata in resolving important issues in economic research and policy. In closing we note that the limit on detail in the LRD is imposed by the establishment collection unit. However, within this limit, available computer technology makes its possible to classify and aggregate the data along a variety of dimensions. No longer does data collection and dissemination need to be tied to only "one" system. In contrast to the past, where tabulations of the data have been restricted to SIC classifications and particular localities, the use of the data can be the determining factor in classification.

This principle has been described recently in work conducted at the Center involving the SIC system.[24] After recounting numerous complaints and shortcomings which have been voiced about the SIC

system, Andrews and Abbott (1988) examined how well it classifies the data under various conceptual frameworks which have been proposed as a basis for the SIC system (markets, production compatibility, etc.). They find the current system is a compromise which satisfies no particular objective. Extensions of the research to show (through the use of cluster algorithms) how the LRD data would look under various classification criteria are currently underway. But, the real message we draw from their work is that the data are sufficiently detailed and rich to support many classifications developed from objectively determined criteria. One such criterion is the grouping of producers based on the closeness of their production technologies as judged by input proportions.[25] There are other possibilities. Regardless of the desired categorizations of the data the Center is attempting to build into the LRD software the flexibility to organize the raw observations according to research needs.

Appendix A.

Variables in the LRD

Symbol	Variable	Availability*
ppn	permanent plant number	
id	identification number	
ind	tabulated industry code	
ppc	primary product class	
pisr	primary industry specialization ratio	
ppsr	primary product specialization ratio	
i13	status of establishment	
tvs	total value of shipments	
ei	employer identification number	
dind	derived industry code	
et	establishment type (0=ASM)	C
ar	administrative record (1=AR)	C
cc	coverage code	
sc	source code	
lfo	legal form of organization	C
st	state code	
smsa	smsa code	
cou	county code	
plac	place code	
va	value added	
vr	value of resales	
rcw	receipts for contract work	
msc	miscellaneous receipts	
te	total employment	
pw1	production workers: march	
pw2	production workers: may	
pw3	production workers: august	
pw4	production workers: november	
pw	production workers (average)	
ph1	manhours: january-march	
ph2	manhours: april-june	
ph3	manhours: july-september	
ph4	manhours: october-december	
ph	total manhours	
sw	total salaries and wages	
ww	wages: production workers	
ow	wages: other employees	
lc	total supplemental labor costs	
le	legally required supplemental labor costs	
vlc	voluntary supplemental labor costs	

* The variable is available for all years and all establishments except as noted: A = collected for ASM establishments only;
C = collected in census years only

Symbol	Variable	Availability*
cp	cost of materials, parts, etc.	
cr	cost of resales	
cf	cost of fuels	
ee	cost of purchased electricity	
pe	quantity purchased electricity	
cw	cost of contract work	
cpc	cost of purchased communications	A 77 & 82
fib	b.o.y. inventory: finished goods	
wib	work-in-progress	
mib	materials	
fie	e.o.y. inventory: finished goods	
wie	work-in-progress	
mie	materials	
tib	b.o.y. inventory: total	
tie	e.o.y. inventory: total	
nb	new building expenditures	
nm	new machinery expenditures	
ue	used capital expenditures	
bab	building assets - b.o.y.	A; after 73
mab	machinery assets - b.o.y.	A; after 73
bae	building assets - e.o.y.	A
mae	machinery assets - e.o.y.	A
br	building rents	A
mr	machinery rents	A
bd	building depreciation	A; after 76
md	machinery depreciation	A; after 76
brt	building retirements	A; after 76
mrt	machinery retirements	A; after 76
rbs	building repair	A; 77 & 82
rm	machinery repair	A; 77 & 82
m	material code	C
mqpc	quantity produced and consumed	C
mqdc	quantity received and consumed	C
mc	delivered cost	C
pi	product code	
pqp	product quantity produced	C
pqs	product quantity shipped	C
pv	product value shipped	
pqit	quantity of interplant transfers	C
pvit	value of interplant transfers	C
pqpc	quantity produced and consumed	C

* The variable is available for all years and all establishments except as noted: A = collected for ASM establishments only;
C = collected in census years only
b.o.y. = beginning of year

e.o.y. = end of year

Appendix B.

Number of Establishments in the LRD for each year:

YEAR	NUMBER OF ESTABLISHMENTS	NUMBER OF ADMINISTRATIVE RECORD CASES
1963	305747	*
1967	305611	118,622
1972	312398	122,158
1973	73460	-
1974	68262	-
1975	71145	-
1976	70346	-
1977	350648	144,648
1978	73853	-
1979	57559	-
1980	55953	-
1981	55045	-
1982	348384	128,307
1983	51619	-
1984	56551	-
1985	55128	-

* There were no administrative record cases in 1963.
\- There are no administrative record cases in the ASM.

REFERENCES

Abbott III, Thomas A. (1988), "Price Dispersion in U.S. Manufacturing," Center for Economic Studies Working Paper, U.S. Bureau of the Census.

Abbott III, Thomas A. and Stephen H. Andrews (1988), "An Examination of the Standard Industrial Classification of Manufacturing Activity using the Longitudinal Research Data Base," Center for Economic Studies Working Paper, U.S. Bureau of the Census.

Abbott III, Thomas A., Robert H. McGuckin, and Paul Herrick, "Advanced Technology Products and the U.S. Trade Balance (Forthcoming).

Bean, Alden, Stephen H. Andrews, and John B. Guerard, "R&D Management and Corporate Financial Policy," Management Science, (forthcoming).

Davis, Steve J. and John Haltiwanger (1987), "Establishment-Specific Labor Demand Disturbances and Unemployment in U.S. Manufacturing Industries," (a research proposal to the Center for Economic Studies).

Dunne, Timothy and Mark J. Roberts (1986), "Measuring Firm Entry, Growth, and Exit with Census of Manufactures Data," mimeo, Pennsylvania State University.

Dunne, Timothy, Mark J. Roberts, and Larry Samuelson (1987), "The Impact of Plant Failure on Employment Growth in the U.S. Manufacturing Sector," mimeo, Pennsylvania State University.

Dunne, Timothy (1988), "Firm Entry and Industry Evolution in the U.S. Manufacturing Sector: Measurement and Analysis," CES Working Paper, U.S. Bureau of the Census.

Gollop, Frank M. and James L. Monahan (1986), "From Homogeneity to Heterogeneity: An Index of Diversification," Center for Economic Studies Working Paper, U.S. Bureau of the Census.

Griliches, Zvi (1984), "Data Problems in Econometrics," NBER Technical Paper No. 39, July.

Guerin-Calvert, Margaret E., Robert H. McGuckin, and Frederick R. Warren-Boulton (1987), "State and Federal Regulation in the Market for Corporate Control," Antitrust Bulletin, Spring 1987.

Hazilla, Michael and Raymond Kopp (1986), "Plant Level Productivity

1972-81: Measurement Using a Large Panel of Manufacturing Establishments," Working Paper.

Kokkelenberg, Edward C. and Sang V. Nguyen (1987), "Forecasting Comparison of Three Flexible Functional Cost Forms," 1987 Proceedings of the Business and Economic Statistics Section, American Statistical Association.

Kokkelenberg, Edward C. and Sang V. Nguyen (1987), "The Stock of Research and Development Knowledge and Multi-Factor Productivity Growth." This paper was presented at the American Economic Association meeting in Chicago, December 27-30, 1987.

Lichtenberg, Frank R. (1987), "The Effects of R&D and Fixed Investment on Productivity," presented at the Allied Social Science Associations Meeting, December 1987.

Lichtenberg, Frank R. and Donald Siegel (1988), "Productivity and Changes in Ownership of Manufacturing Plants," Center for Economic Studies Working Paper, U.S. Bureau of the Census.

Lichtenberg, Frank R. and Zvi Griliches (August 1986), "Errors of Measurement in Output Deflators," NBER Working Paper Series #2000.

McGuckin, Robert H. and Stephen H. Andrews (1988), "The Performance of Lines of Business Purchased in Conglomerate Acquisitions." This paper was presented at the American Economic Association meeting in Chicago, December 27-30, 1987.

McGuckin, Robert H. and James L. Monahan (1987), "High Technology Goods and the U.S. Trade Deficit," U.S. Bureau of the Census Internal Report.

McGuckin, Robert H. and Sang V. Nguyen (1988), "Use of 'Surrogate' Files to Conduct Economic Studies with Longitudinal Microdata," presented at the U.S. Bureau of the Census Fourth Annual Research Conference, March 1988.

McGuckin, Robert H., Frederick R. Warren-Boulton, and Peter Waldstein (1988), "Analysis of Mergers Using Stock Market Returns," U.S. Department of Justice, Antitrust Division, Economic Analysis Group Discussion Paper #EAG 88-1.

McGuckin, Robert H. and Peter Zadrozny (1987), "Long Run Expectations and Capacity," Center for Economic Studies Working Paper, U.S. Bureau of the Census.

Ravenscraft, David J. and F.M. Scherer (1987), Mergers, Sell-Offs, and Economic Efficiency (Washington, D.C.: Brookings Institution).

Solow, John L. (1987), "The Capital-Energy Complementarity Debate Revisited," American Economic Review 77:605-614.
The LED Technical Documentation (1987). The Center for Economic Studies.

ENDNOTES

1. Concentration ratios for 4-digit industries and 5-digit product classes were developed in the early sixties in response to antitrust policy concerns about the problem of discerning the extent to which monopoly power characterized U.S. manufacturing. The theory that a tight oligopolistic structure was a necessary (but not sufficient) condition for monopoly pricing guided the development of special reports on concentration by the Census Bureau. While it is now well accepted that the information collected by the Census Bureau was not sufficient to distinguish between efficiency and monopoly power as explanations of high profit margins, research on these data provided understanding regarding the links between firm behavior and performance.

In contrast to the first example which involved the retabulation of data already being collected for other purposes, another more recent example involved the development of a new survey of business owners. This survey was designed to collect and link information on the individual characteristics of business owners (for example, minority status, education, etc.) to the type and operational aspects of the particular businesses they own. The survey was designed by economists and other social science researchers interested in the nature of "entrepreneurship."

2. Abbott (1988) shows that the use of aggregate industry price deflators leads to biased estimates of productivity growth and production functions estimated in first differences. Siegel and Lichtenberg (1987) found that failure to account for the diversified structure of a firm's production when applying price deflators has a substantial effect on estimates of the role of technical change in total factor productivity. Similar findings are also reported by Nguyen and Kokkelenberg (1987). Finally, in a recent theoretical paper, using examples from the Census Bureau's Survey of Plant Capacity and earlier work performed under Center sponsorship, McGuckin and Zadrozny (1988) describe several econometric problems with existing work on capacity utilization most of which employs aggregate data.

3. A comparison of trade balances derived from allocating aggregate industries to high-tech and low-tech categories with those derived by aggregating information on individual products separated into high-tech and low-tech categories showed substantial level and trend differences. See McGuckin and Monahan (1987) and Abbott, McGuckin and Herrick (1988).

4. The Center's limited computer resources, coupled with the size of its databases, precludes using relational data structures which, in effect, allow all data items to act as keys.

5. A series of papers dealing with birth, growth and death of business entities resulted from one of the early research efforts at the Center. See Roberts, Samuelson and Dunne (1986), (1987), and (1988). Similarly, work by Gollop and Monahan (1986) and (1987) used the full unbalanced LED panel. Neither of these efforts yielded general software applicable to other projects.

6. We note that the change of the database name from LED to LRD was done for three reasons. First, the name emphasizes the new database structure used for updating and extracting microdata. Second, the name focuses attention to the primary use of the data: research and analysis. Finally, the name change is intended to eliminate confusion which developed because the Time-Series and LED files became synonymous in the minds of many people. This was particularly troublesome because the time series was a restricted balanced panel.

7. To this end, the Center is developing software which will enable a researcher to select alternative estimation strategies.

8. In this section, we focus on the size of the reporting unit in determining its probability of selection. In practice the sampling design is more complex, including factors such as existence of the unit in the previous panel and industry affiliation. In the past, location may also have been included in the sample design. It is not currently a criterion variable.

9. We are unsure of the exact effect of this change. There still appears to be 3,000 to 4,000 fewer complete firms than would be obtained under the pre-1979 procedure.

10. Standard Metropolitan Statistical Area

11. At this juncture the Center has not done much analysis with the geographic codes except in the area of state tabulations.

12. Current Industrial Reports data are not linked to the LRD. These reports contain yearly and sometimes monthly unit value data for many detailed SIC classifications. The Center hopes eventually to link these data to the LRD.

13. An exception to classification of plants by primary industry is the Census product concentration ratios. Unfortunately budget reductions may lead to elimination of this and other reports based on product level detail. Such detail requires additional edit checks by industry analysts. The only Center based research to focus on decision units classified by product is reported in the series of papers by Roberts, Samuelson and Dunne (1987). As

noted in an earlier footnote they examined entry, exit and survival patterns for firms classified by product class. This methodology has the advantage of looking at resource movements among groups which appear closer (at least from the demand side) to markets than those obtained using industry groupings. Unfortunately, other information, such as price-cost margins cannot be linked to the resource movements.

14. See Lichtenberg and Siegel (1988), and Hazilla and Kopp (1987).

15. The proposal for extended development of a plant level database based on the experience gained in the Phase I project is pending at NSF. If funding for an extension of the RFF's research is obtained they will begin to develop an unbalanced panel containing information on each manufacturing establishment employing over 250 employees and a random selection of smaller establishments from ASM panels. The total number of these establishments will exceed 35,000 in a continuous panel and over 60,000 when new plants are added to form the discontinuous panel. The length of the panel's time dimension will continue to grow as new ASM and CM become available. The RFF's research will go beyond extraction of data from Census Bureau files; this has already been undertaken by the Center and is available in the LRD. Rather, RFF's purpose is to combine the LRD with additional information not contained in the CM or ASM and to employ economic theory to construct a class of variables that fully enumerates the prices paid and quantities employed of all productive factors and manufactured products. The data set would include: the net stocks of structures, equipment and inventories and the implicit rental prices for each, employment information on production and nonproduction labor, the prices and quantities for seven forms of energy, up to 30 purchased intermediate inputs and an equal number of multiple outputs. In short, the RFF research will provide a new panel data set consisting of new data variables combined with edits, imputations and transformations of the data contained in the LRD.

16. A recent paper by Griliches and Lichtenberg (1986) discusses these differences.

17. See, for example, Scherer and Ravenscraft (1987) which uses accounting data and McGuckin, Warren-Boulton and Waldstein (1988), and Guerin-Calvert, McGuckin, and Warren-Boulton (1987) which reports premiums based on financial market data.

18. Lichtenberg (1987) and Bean (1988).

19. A related area of future work is the development of longitudinal panels for census programs conducted outside manufacturing. Such a program is already underway for the agriculture census.

20. See Davis and Haltiwanger (1987).

21. Precise criteria for evaluating disclosure risk in economic microdata like those found in the LRD are not yet available. Masked microdata files pertaining to demographic data have been released by the Census Bureau. These data sets involve samples of 100,000 individuals or more. The skewed size distribution of establishments and the relatively small numbers in the population make such large samples impossible for LRD data. The Center has begun work on creation of public use microdata files, but the problem is very difficult. See McGuckin and Nguyen (1988).

22. See McGuckin and Nguyen (1988) for an extended discussion and several proposals.

23. In some cases, for projects involving data tabulations, arrangements can be made for the Center staff to undertake the data work directly.

24. See Andrews and Abbott (1988).

25. This type of procedure was used by Gollop and Monahan (1986) in constructing an index of diversification. They measured the closeness of products by the technologies of pure producers.